基于改进 CFSFDP 算法的文本聚类方法及其应用*

詹春霞　王荣波　黄孝喜　谌志群
(杭州电子科技大学计算机学院　杭州 310018)

摘要:【目的】针对 CFSFDP(Clustering by Fast Search and Find of Density Peaks)算法利用局部密度和距离的乘积选择聚类中心而导致聚类结果不理想的问题进行改进。【方法】提出一种基于粒子群算法的 CFSFDP 算法, 通过粒子群算法寻找 CFSFDP 算法中的最佳局部密度和距离阈值, 得到相对较高的局部密度和距离的聚类中心, 减少离散点对数据中心选取的影响, 并在某高考咨询平台提供的考生问题库中随机选取数据集进行试验。【结果】实验结果表明, 在不同的数据集中, 本文算法相对于基本的 CFSFDP 算法在准确率、召回率、F 值上均有明显提高。【局限】文本处理时没有考虑语义关系。【结论】本文方法有很好的聚类效果, 应用在高考咨询库中能够有效地减轻被咨询方的工作量并且帮助快速回答考生的问题。

关键词: CFSDFP　聚类中心　粒子优化群算法
分类号: TP391

1 引　言

随着信息时代的到来, 互联网中的数据以爆炸式增长, 如何从这些海量的数据中获取有用的信息并对数据进行有效的处理分析是当前研究的热点。数据挖掘[1]中的一个热门分支就是聚类[2], 它是一种无监督学习方法, 无需任何先验知识, 按照某种相似性度量方式, 找到数据之间的共性, 将数据集划分成若干个不同的类。划分到同一个类中的数据相似度高、差异小, 而不同类之间的数据相似性较低。迄今为止, 对聚类方法的研究已经长达几十年, 它在医学、模式识别、图像处理、用户兴趣推荐等方面具有广泛的应用, 推动了社会的发展, 改善了人们的生活。

目前, 聚类算法主要分为 5 大类[2-3]: 基于层次的方法、基于划分的方法、基于密度的方法、基于模型的方法和基于网络的方法。每一类聚类方法都有一些经典算法[3], 在文本处理方面有着广泛的应用。但是鉴于数据的多样性和复杂性, 没有任何一种聚类算法可以普遍适用于各种数据集, 每一类方法都有各自的优点和缺陷, 不同的聚类算法会得到不同的聚类结果。本文对比实验中, Agglomerative Clustering 算法和 DBSCAN 算法分别是基于层次和基于密度的方法, 基本的 CFSFDP 算法是由 Rodriguez 和 Laio 提出的一种新的密度聚类算法[4], 该算法具有能够发现任意形状的数据集且快速简单的优点。张文开进行了基于密度的层次聚类算法研究[5], Mehmood 等进行了基于 CFSFDP 算法的模糊聚类研究[6], 马春来等提出一种基于簇中心点自动选择策略的密度峰值聚类算法[7]。由于 CFSFDP 算法聚类中心的选取取决于数据点密度和距离乘积的大小, 乘积越大越有可能是聚类中心, 而数据集中密度大距离小或距离大密度小的数据点之间的乘积也可能很大而被误认为是聚类中心。因此本

通讯作者: 詹春霞, ORCID: 0000-0001-8790-5520, E-mail: 1239350526@qq.com。

*本文系国家自然科学基金青年基金项目“引入涉身认知机制的汉语隐喻计算模型及其实现”(项目编号:61103101)、国家自然科学基金青年基金项目“基于马尔科夫树与 DRT 的汉语句群自动划分算法研究”(项目编号: 61202281)和教育部人文社会科学研究青年基金项目“面向信息处理的汉语隐喻计算研究”(项目编号: 10YJCZH052)的研究成果之一。

文通过引入粒子群算法找到一对密度距离阈值，数据集中密度和距离均大于这对阈值的数据点为数据中心，减少了离散点对聚类中心选取的影响，实现了聚类中心的自动选择，减少了人工干预的过程。

本文的实验数据来自于某高考咨询平台自动问答APP，其中的数据都是学生对于所报考大学的录取情况、学校基本信息等方面的问题，对其中文本的聚类有利于完善机器人知识库，提高对学生问答的准确率。将该算法应用于从中抽取的数据集中，证明了本文聚类算法的有效性。

2 相关工作

2.1 CFSFDP 算法

CFSFDP[4]聚类算法的基本思想是：首先计算数据点的密度及距离，其次选取聚类中心，最后对非聚类中心点进行归类操作。其中，对聚类中心的选取是该算法的关键。聚类中心点具有两个重要的特征：聚类中心本身密度比较大，它是由一些密度比它小的数据点包围；与其他比其密度高的数据点之间的距离都比较大。基本的 CFSFDP 算法选取聚类中心的方法具有很大的缺点：数据集中密度大距离小或密度小距离大的数据点乘积也可能很大而被误认为是聚类中心；聚类中心的个数无法自动确定，需要一个人工干预的过程。

(1) 局部密度和距离

设有数据集 $s=\{x_i\}_{i=1}^{N}$，$I_s=\{1,2,\cdots,N\}$，$d_{i,j}$ 表示数据点 x_i 和数据点 x_j 之间的距离。对于数据集 s 中的每一个数据点 x_i，可以用两个变量进行刻画：局部密度和距离。计算局部密度 ρ_i，如公式(1)所示[4]。

$$\rho_i=\sum_{j\in I_s\{i\}}\varphi(d_{i,j}-d_c) \qquad (1)$$

其中，$\varphi(x)=\begin{cases}1, x<0\\0, x\geqslant 0\end{cases}$，参数 d_c>0 为截断距离。由公式(1)可知，每个数据点的密度是在数据集 s 中与该数据点的距离小于 d_c 的数据点个数(不包括本身)。

当数据点 x_i 具有最大的局部密度时，其距离为 s 与 x_i 距离最大的数据点与 x_i 之间的距离。除此之外，对于其他不具有最大密度的数据点，距离表示在所有局部密度大于 x_i 的数据点当中，与 x_i 的距离最小的数据点与 x_i 之间的距离。其计算如公式(2)所示[4]。

$$\delta_{qi}=\begin{cases}\min\limits_{j<i}\{d_{qi,qj}\} & i\geqslant 2\\ \max\limits_{j\geqslant 2}\{d_{q1,qj}\} & i=1\end{cases} \qquad (2)$$

其中，$\rho_{q1}\geqslant\rho_{q2}\geqslant\cdots\geqslant\rho_{qN}$。

(2) 决策图

以局部密度ρ为横轴，距离δ为纵轴，对数据点的局部密度和距离刻画出相应的决策图。图1是由28个数据点包含的散点图，相应的决策图如图2[4]所示。

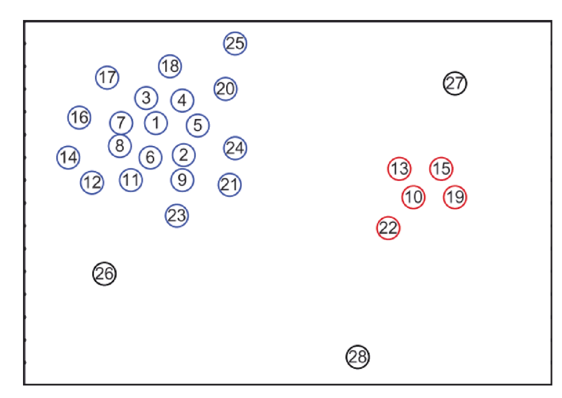


图 1 散点图[4]

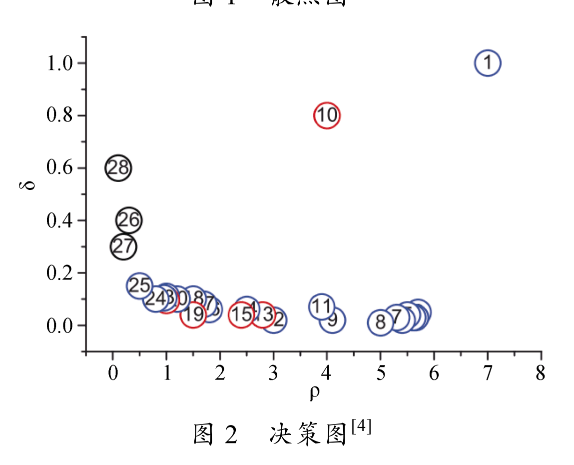


图 2 决策图[4]

2.2 粒子群算法

粒子群算法[8-9]中，种群中的粒子都有决定自身方向和位置的速度和由适应度函数决定的适应度值，每个粒子通过向自身和群体曾达到的最优位置靠拢来动态调节自身位置，通过迭代得到最优解。

假设粒子群的种群规模为 N，种群中个体维度为 D，每一个粒子都有两个属性：当前的位置x_i和飞行的速度 v_i，可表示为 $x_i=(x_{i,1},x_{i,2},\cdots,x_{i,D})$，$v_i=(v_{i,1},v_{i,2},\cdots,v_{i,D})$，其中 i=1,2,⋯N。$P_i$ 为粒子 x_i 在搜索解的过程中适应度值最高的位置；P_g 为整个粒子群中

粒子所达到的最优位置，即 P_g 是所有 P_i 当中的适应度最大的值。在每一次迭代的过程中，每个粒子通过这两个极值来调整自身的位置和飞行速度。位置和速度的更新如公式(3)和公式(4)所示[9]。

$$x_i = x_i + v_i \quad (3)$$

$$v_i = w \times v_i + c_1 \times r_1 \times (P_i - x_i) + c_2 \times r_2 (P_g - x_i) \quad (4)$$

其中，c_1、c_2 为正数，称之为加速因子；r_1、r_2 为[0,1]中均匀分布的随机数；w 是惯性权重因子。v_{max} 是粒子的最大速率，$v_i \in [-v_{\max}, v_{\max}]$，当粒子的飞行速度超过 v_{max} 时，粒子的飞行速度即为 v_{max}。

3 基于改进 CFSFDP 算法进行文本聚类

由于基本的 CFSFDP 算法存在上述缺陷，本文引入了粒子群算法。改进 CFSFDP 算法的主要思想为：利用粒子群算法调节 CFSFDP 算法中聚类中心的选取。即通过粒子群算法得到一个密度和距离的阈值，在 CFSFDP 算法中密度值和距离均大于这个阈值的数据点为聚类中心，根据选取出来的聚类中心进行聚类，根据聚类结果计算适应度值，将其作为粒子群算法更新的判断依据。将其运用到文本聚类中，通过计算文本之间的相似性，计算出每条文本的密度和局部距离，实现文本聚类。算法的流程如图 3 所示。

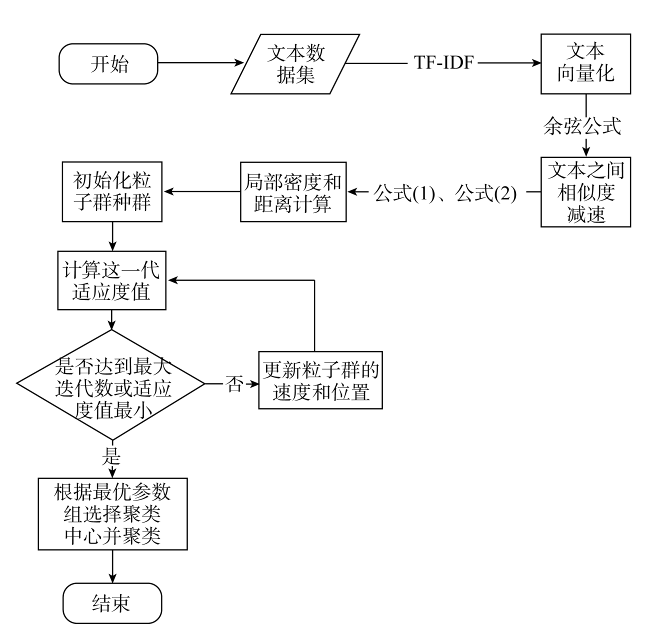


图 3　算法流程

对于全部的文本数据集，采用目前最广泛的文本处理方法，基于 TF-IDF(Term Frequency-Inverse Document Frequency)[10-12]的向量空间模型 VSM (Vector Space Model)[12]来表示文本，向量的每一维为对应的特征词在该文本中的权重。文本相似度的度量方式是依据余弦距离[13-14]，值越大，两条文本之间越相似，距离越近。

粒子群算法在迭代过程中以适应度函数为依据，适应度值越高说明该粒子的适应能力越好，则会对下一代粒子的进化产生影响，从而产生最优解。本文采用 Rand 系数[15]的倒数作为粒子群算法的适应度函数，以此来评价聚类结果的好坏。

算法的具体过程如下：

①对文本集 S 进行预处理，计算每条文本特征词的权重，得到每条文本的向量化表示并且计算文本之间的相似度。

②将①中得到的每个向量作为一个数据点，根据公式(1)和公式(2)计算每个数据点的局部密度和距离。

③初始化粒子群算法的参数，主要包括种群规模 m、惯性权重 w、学习因子 c_1 和 c_2、最大迭代次数 t 等。随机生成种群的初始速度和初始位置，将粒子的初始位置赋值给粒子的初始最优位置 P_i，根据各个粒子的最优位置找到全局最优位置 P_g，位置由密度和距离确定。

④计算每个粒子对应的聚类结果。将每个粒子的位置传递给 CFSFDP 算法，数据集中局部密度和距离均大于此粒子位置的数据点记为聚类中心，并使用数据点归属方法对非聚类中心点进行归属，完成聚类操作。

⑤对于每个粒子对应的聚类结果计算适应度值，更新当前每个粒子的最优位置。并且根据每个粒子的最优位置更新种群全局最优位置，更新每个粒子的位置和速度。

⑥判断是否满足收敛条件或是否达到最大迭代次数，若是，返回⑦；否则，迭代次数加 1 后执行④。

⑦根据种群的全局最优位置选取聚类中心，通过数据点的归属方法完成聚类，得到最终文本集的聚类结果，算法结束。

4 实验结果及分析

4.1 数据集

实验中的数据来自于从高考咨询平台 APP 中考生向学校招生办提出的问题，从问题库中随机选取 7 个类别：询问学校代码和专业代码类、军训有关事项、高考加分情况、分数极差情况、询问招生办电话、省控线有关信息、是否有退档情况。从每一类中随机选取构造包含 7 类数据的数据集，共构造出 data1050、data3100、data5000 三个数据集，分别包含 1 050、

3 100、5 000 条数据，其中每个数据集中包含的各个类数如表 1 所示。利用"结巴分词"、停用词处理等对数据集进行预处理，并且对不同的数据集进行实验分析比较。本文通过纯度(Accuracy)、精度(Precision)、召回率(Recall)和 F 度量值(F-Measure)[16-17] 4 个评价指标衡量聚类效果。

表 1　文本数据集

数据集 \ 类	代码	军训	加分	极差	电话	省控线	退档
data1050	200	100	200	100	150	200	100
data3100	600	300	600	300	500	500	300
data5000	1 000	400	1000	400	900	900	400

4.2　实验结果分析

分别用层次聚类算法(Agglomerative Cluster)[18-19]、DBSCAN 算法[20]、基本的 CFSFDP 算法、以及本文算法对抽取的三个数据集进行聚类比较。其中 Agglomerative Clustering 算法因其可以适用于任意形状和任意属性的数据集在文本聚类方面也有广泛的应用。Agglomerative Clustering 算法采用在三个数据集中设定的类别数目实验效果最佳的 7。粒子群算法中种群数量设定为 50，最大迭代次数为 30，加速因子为 2，惯性权重因子为 0.5。DBSCAN 算法是基于密度的聚类算法中的一种经典算法，具有较强的代表性。针对 DBSCAN 算法进行多次实验，在数据集 data1050 中选取参数 eps=0.8、minPts=30，数据集 data3100 中参数 eps=0.8、minPts=70，数据集 data5000 中选取参数 eps=0.8、minPts=110 效果相对最佳的实验结果。层次聚类算法、DBSCAN 算法、基本的 CFSFDP 算法、本文算法总体 F 值比较如图 4 所示。

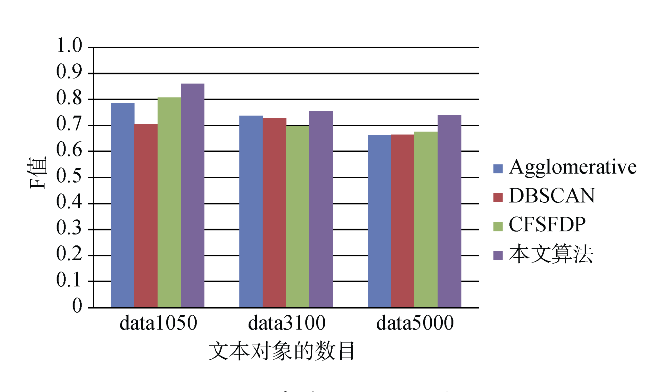


图 4　聚类效果的比较

可见本文算法相对于其他三个算法在不同的数据集中聚类效果都较好。实验结果如表 2 所示，可以看出本文算法在高考问询文本库中相比其他三个算法都有较好的效果。其中基本的 CFSFDP 算法由于噪声点的影响造成同一个类中两个及以上的数据点成为数据中心导致聚类算法的不准确性。DBSCAN 算法由于对参数 eps(被认为是同一个类的文本之间的最大距离)和 minPts(一条文本与其他文本的距离小于 eps 的个数大于等于 minPts 视为聚类中心)特别敏感，在类中的数据发布密度不均匀的时候，eps 较小时，密度小的类会被划分成多个相似的类，eps 较大时，会使得距离较近且密度较大的类被合并成一个较大的类，导致聚类效果不理想。层次算法比 DBSCAN 算法具有更好的效果，但是 Agglomerative Clustering 算法的计算复杂度太高。

表 2　4 种算法的 Accuracy、Precision、Recall、F-Measure 值比较

算法	数据集	Accuracy	Precision	Recall	F-Measure
Agglomerative	data1050	0.7305	0.7743	0.7969	0.7854
	data3100	0.7077	0.6976	0.7811	0.7370
	data5000	0.6808	0.6598	0.6627	0.6612
DBSCAN	data1050	0.6486	0.6795	0.7332	0.7052
	data3100	0.6797	0.6761	0.7880	0.7278
	data5000	0.6006	0.6270	0.6500	0.6643
CFSFDP	data1050	0.8171	0.8050	0.8090	0.8070
	data3100	0.750	0.7375	0.6617	0.6975
	data5000	0.7425	0.7438	0.6189	0.6756
本文算法	data1050	0.8333	0.7171	0.9098	0.8609
	data3100	0.7574	0.7421	0.7676	0.7546
	data5000	0.7712	0.7340	0.7450	0.7395

5　结　语

本文针对 CFSFDP 算法聚类中心选取的武断性的问题，提出一种基于粒子群算法的 CFSFDP 算法。引入粒子群算法找到一对阈值，将大于这对阈值的数据点作为聚类中心，减少离散点对聚类结果的影响，提高了聚类准确性。将此算法应用在从某高考咨询平台问题库中随机提取的问题中，验证了本文算法的有效性和准确性，能够帮助考生更准确高效地获得答案并且减轻了被咨询方的咨询量，大大节省了双方的时

间。但是该算法也存在局限性，由于粒子群本身算法的特性，在计算高纬度的问题时，粒子群优化算法需要的粒子数较多，导致计算复杂度通常很高。

参考文献：

[1] Tan P N, Steinbach M, Kuma V. Introduce to Data Mining [M]. Addison-Wesley Professional, 1988.

[2] 孙吉贵，刘杰，赵连宇．聚类算法研究[J]．软件学报，2008, 19(1): 48-61. (Sun Jigui, Liu Jie, Zhao Lianyu. Clustering Algorithms Research[J]. Journal of Software, 2008, 19(1): 48-61.)

[3] 史梦洁．文本聚类算法综述[J]．现代计算机，2014(2): 3-6. (Shi Mengjie. Summary of Text Clustering Algorithms[J]. Modern Computer, 2014(2): 3-6.)

[4] Rodriguez A, Laio A. Clustering by Fast Search and Find of Density Peaks[J]. Science, 2014, 344(6191): 1492-1496.

[5] 张文开．基于密度的层次聚类算法研究[D]．合肥：中国科学技术大学，2015. (Zhang Wenkai. Research on Density-based Hierarchical Clustering Algorithm[D]. Hefei: University of Science and Technology of China, 2015.)

[6] Mehmood R, Bie R, Dawood H, et al. Fuzzy Clustering by Fast Search and Find of Density Peaks[C]//Proceedings of the 2015 International Conference on Identification, Information, and Knowledge in the Internet of Things. 2015.

[7] 马春来，单洪，马涛．一种基于簇中心点自动选择策略的密度峰值聚类算法[J]．计算机科学，2016, 43(7): 255-258. (Ma Chunlai, Shan Hong, Ma Tao. Improved Density Peaks Based Clustering Algorithm with Strategy Choosing Cluster Center Automatically[J]. Computer Science, 2016, 43(7): 255-258.)

[8] Kennedy J, Eberhart R. Partical Swarm Optimization[C]// Proceeding of the 1995 IEEE International Conference on Neural Networks. 1995.

[9] 刘建华．粒子群算法的基本理论及其改进研究[D]．长沙：中南大学, 2009. (Liu Jianhua. The Basic Theory of Partical Swarm Optimization and Its Improvement[D]. Changsha: Central South University,2009.)

[10] 黄承慧，印鉴，侯昉．一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]．计算机学报，2011, 34(5): 856-864. (Huang Chenghui, Yin Jian, Hou Fang. A Text Similarity Measurement Combining Word Semantic Information with TF-IDF Method [J]. Chinese Journal of Computer, 2011, 34(5): 856-864.)

[11] Aizawa A. An Information-treoretic Perspective of TF-IDF Measures[J]. Information Processing and Management, 2003, 39(1): 45-65.

[12] Salton G, Buckley C. Term Weight Approaches in Automatic Text Retrieval [J]. Information Processing and Management, 1988, 24(5): 513-523.

[13] 谭静．基于向量空间模型的文本相似度算法研究[D]．成都：西南石油大学，2015. (Tan Jing. Research on Text Similarity Algorithm Based on Vector Space Modal[D]. Chengdu: Southwest Petroleum University, 2015.)

[14] 赵俊杰，胡学钢．基于文本分类的文档相似度计算[J].微型电脑应用，2008, 24(12): 46-47. (Zhao Junjie, Hu Xuegang. Simility Calculation Based on Text Classification [J]. Microcomputer Application, 2008, 24(12): 46-47.)

[15] Halkidi M, Batistakis Y, Vazirgiannis M. On Clustering Validation Techniques[J]. Journal of Intelligent Information Systems, 2015, 17(2-3): 107-145.

[16] Liang J, Bai L, Dang C, et al. The K-Means-Type Algorithms Versus Imbalanced Data Distributions[J]. IEEE Transactions on Fuzzy Systems, 2012, 20(4): 728-745.

[17] 张鸣．符号数据聚类评价指标研究[D]．太原：山西大学，2013. (Zhang Ming. Study on the Evaluation Index Symbol of Data Clustering[D]. Taiyuan: University of Shanxi, 2013.)

[18] Franti P, Virmajoki O, Hautamaki V. Fast Agglomerative Clustering Using a K-nearest Neighbor Graph [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2006, 28(11): 1875-1881.

[19] 段明秀．层次聚类算法的研究及应用[D]．长沙:中南大学，2009. (Duan Mingxiu.Research and Application of Hierarchical Clustering Algorithm[J]. Changsha: Central South University, 2009.)

[20] 冯少荣，肖文俊．DBSCAN 聚类算法的研究与改进[J].中国矿业大学学报，2008, 37(1): 106-111. (Feng Shaorong, Xiao Wenjun. An Improved DBSCAN Clustering Algorithm[J]. Journal of China University of Mining & Technology, 2008, 37(1): 106-111.)

作者贡献声明：

詹春霞：提出研究思路，采集分析数据；

詹春霞，王荣波：进行实验，起草并修改论文；

王荣波，黄孝喜，谌志群：修改论文。

利益冲突声明：

本文实验中所用的数据由达言公司提供，实验数据仅限于科学研究，不可在网络传播或用于其他用途。

支撑数据：

支撑数据由作者自存储，http://pan.baidu.com/s/1bpL0WcJ。

[1] 王荣波. data1050.rar. 数据集 data1050 中包含的数据.
[2] 王荣波. data3100.rar. 数据集 data3100 中包含的数据.
[3] 王荣波. data5000.rar. 数据集 data5000 中包含的数据.

收稿日期: 2016-12-30
收修改稿日期: 2017-03-15

Application of Text Clustering Method Based on Improved CFSFDP Algorithm

Zhan Chunxia　Wang Rongbo　Huang Xiaoxi　Chen Zhiqun
(School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: [**Objective**] This paper aims to improve the un-satisfactory performance of CFSFDP (clustering by fast search and find of density peaks) algorithm with the help of based on particle swarm optimization. [**Methods**] First, we determined the cluster centers by searching optimal local density and distance thresholds to increase the accuracy of results. These clustering centers have relatively high local density and distance, which reduced the influence of discrete points. Then, we examined the proposed method on a randomly selected dataset from the question-answer database of a college entrance exam consulting platform. [**Results**] The modified CFSFDP algorithm had better performance than the original one. [**Limitations**] We did not include the semantic relations to process the texts. [**Conclusions**] The proposed algorithm could achieve good clustering results, and improve the efficiency of the consulting personnel .

Keywords: CFSDFP　Cluster Centers　Particle Swarm Optimization Algorithm